

Extended T-process Regression Models

Zhanfeng Wang^{a,c}, Jian Qing Shi^b, Youngjo Lee^{c,*}

^a*Department of Statistics and Finance, University of Science and Technology of China, Hefei, China*

^b*School of Mathematics and Statistics, Newcastle University, Newcastle, UK*

^c*Department of Statistics, Seoul National University, Seoul, Korea*

Abstract

Gaussian process regression (GPR) model has been widely used to fit data when the regression function is unknown and its nice properties have been well established. In this article, we introduce an extended t-process regression (eTPR) model, which gives a robust best linear unbiased predictor (BLUP). Owing to its succinct construction, it inherits many attractive properties from the GPR model, such as having closed forms of marginal and predictive distributions to give an explicit form for robust BLUP procedures, and easy to cope with large dimensional covariates with an efficient implementation by slightly modifying existing BLUP procedures. Properties of the robust BLUP are studied. Simulation studies and real data applications show that the eTPR model gives a robust fit in the presence of outliers in both input and output spaces and has a good performance in prediction, compared with the GPR and locally weighted scatterplot smoothing (LOESS) methods.

Keywords: Gaussian process regression, selective shrinkage, robustness, extended t process regression, functional data

1. Introduction

Consider a functional regression model

$$y_i = f_0(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \dots, n \quad (1)$$

where $f_0(\mathbf{x}_i)$ is the value of unknown function $f_0(\cdot)$ at the $p \times 1$ observed covariate $\mathbf{x}_i \in \mathcal{X} = R^p$ and ϵ_i is an error term. To fit an unknown function

*Corresponding author. Email: youngjo@snu.ac.kr

f_0 , we may consider a process regression model

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x}), \quad (2)$$

where $f(\mathbf{x})$ is a random function and $\epsilon(\mathbf{x})$ is an error process for $\mathbf{x} \in \mathcal{X}$. A GPR model assumes a Gaussian process (GP) for the random function $f(\cdot)$. It has been widely used to fit data when the regression function is unknown: for detailed descriptions see Rasmussen and Williams (2006), and Shi and Choi (2011) and references therein. GPR has many good features, for example, it can model nonlinear relationship nonparametrically between a response and a set of large dimensional covariates with efficient implementation procedure. In this paper we introduce an eTPR model and investigate advantages in using an extended t-process (ETP).

BLUP procedures in linear mixed model are widely used (Robinson, 1991) and extended to Poisson-gamma models (Lee and Nelder, 1996) and Tweedie models (Ma and Jorgensen, 2007). Efficient BLUP algorithms have been developed for genetics data (Zhou and Stephens, 2012) and spatial data (Dutta and Mondal, 2015). In this paper, we show that BLUP procedures can be extended to GPR models. However, GPR fits are susceptible to outliers in output space (y_i). LOESS (Cleveland and Devlin, 1988) has been developed for a robust fit against such outliers. However, it requires fairly large densely sampled data set to produce good models and does not produce a regression function that is easily represented by a mathematical formula. For models with many covariates, it is inevitable to have sparsely sampled regions. Wauthier and Jordan (2010) showed that the GPR model tends to give an overfit of data points in the sparsely sampled regions (outliers in the input space, \mathbf{x}_i). Thus, it is important to develop a method which produces good fits for sparsely sampled regions as well as densely sampled regions. Wauthier and Jordan (2010) proposed to use a heavy-tailed process. However, their copula method does not lead to a close form for prediction of $f(\mathbf{x})$. As an alternative to generate a heavy-tailed process, various forms of student t -process have been developed: see for example Yu *et al.* (2007), Zhang and Yeung (2010), Archambeau and Bach (2010) and Xu *et al.* (2011). However, Shah *et al.* (2014) noted that the t -distribution is not closed under addition to maintain nice properties in Gaussian models.

In this paper, we develop a specific eTPR model which is closed under addition to retain many favorable properties of GPR models. Due to its special structure of construction, the resulting eTPR model gives computationally efficient algorithm, i.e. a slight modification of the existing BLUP

algorithm provides the robust BLUP procedure. Under the proposed eTPR model, marginal and predictive distributions are in closed forms. Furthermore, it gives a robust BLUP procedure against outliers in both input and output spaces. Properties of the robust BLUP procedure are investigated.

The remainder of the paper is as follows. Section 2 presents an ETP and its properties. Section 3 proposes an eTPR model and discusses the inference and implementation procedures. Robustness properties and information consistency of robust BLUP predictions are shown in Section 3. Numerical studies and real examples are in Section 4, followed by concluding remarks in Section 5. All the proofs are in Appendix.

2. Extended t -process

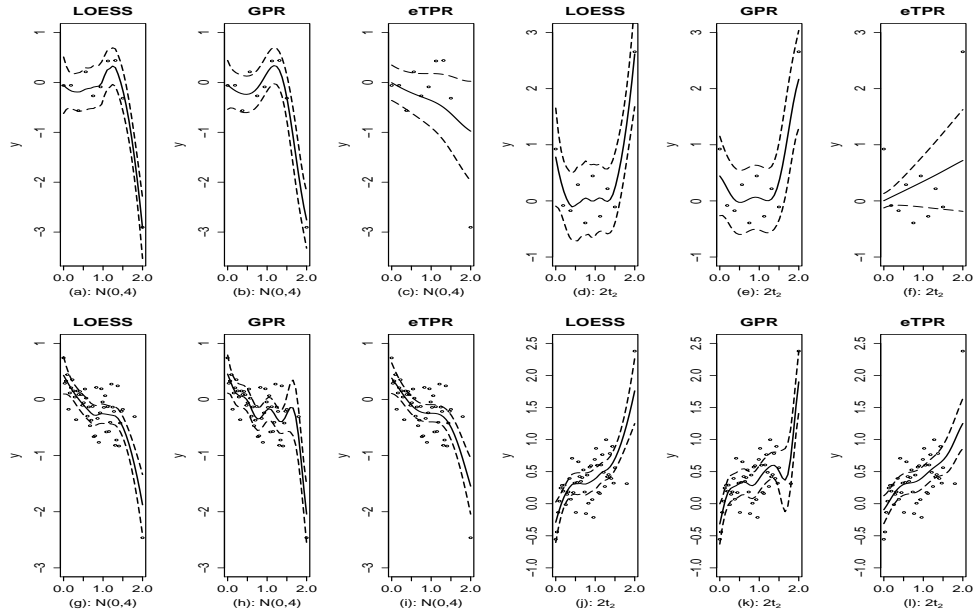


Figure 1: Predictions in the presence of outliers at data point point 2.0 disturbed by additional errors with the normal distribution $N(0,4)$ or the t distribution $2t_2$ where circles represent data points, solid and dashed lines stand for predicted curves and their 95% confidence bounds from the LOESS, GPR and eTPR methods, respectively.

As a motivating example, we generated two data sets with sample sizes of $n = 10$ and $n = 50$ where x_i 's are evenly spaced in $[0, 1.5]$ for the 9 (or

48) data points and the remaining point is at 2.0 (or two points at 1.8 and 2.0). Thus, the remaining point or the two points are sparse ones, meaning they are far away from the other data points in input space. In addition, we also make the data point 2.0 to be an outlier in output space by adding an extra error from either $N(0, 4)$ or $2t_2$, where t_2 is the student t distribution with two degree of freedom. Prediction curves for simulated data are plotted in Figure 1, where circles represent data points, solid and dashed lines stand for prediction and their 95% confidence bounds. The true function is zero. For a small sample size $n = 10$, Figure 1(a-f) shows that LOESS and GPR predictions are similar and the eTPR prediction is the smoothest and shrinks the data point 2.0 the most heavily, i.e. selective shrinkage occurs. For a moderate sample size $n = 50$, Figure 1(g-l) shows that LOESS and eTPR predictions are similar. However, the eTPR prediction still shrinks the most at 2.0. Even though unreported, for a large sample size $n = 100$, all give similar predictions.

Denote observed data set by $\mathcal{D}_n = \{\mathbf{X}_n, \mathbf{y}_n\}$ where $\mathbf{y}_n = (y_1, \dots, y_n)^T$ and $\mathbf{X}_n = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$. For random component $f(\mathbf{u})$ at a new point $\mathbf{u} \in \mathcal{X}$, the best unbiased predictor is $E(f(\mathbf{u})|\mathcal{D}_n)$. It is called a BLUP if it is linear in \mathbf{y}_n . Its standard error can be estimated with $\text{Var}(f(\mathbf{u})|\mathcal{D}_n)$. To have an efficient implementation procedure, it is useful to have explicit forms for the predictive distribution $p(f(\mathbf{u})|\mathcal{D}_n)$, $E(f(\mathbf{u})|\mathcal{D}_n)$ and $\text{Var}(f(\mathbf{u})|\mathcal{D}_n)$.

Let f be a real-valued random function such that $f : \mathcal{X} \rightarrow R$. Analogous to double hierarchical generalized linear models (Lee and Nelder, 2006), we consider a following hierarchical process,

$$f|r \sim GP(h, rk), \quad r \sim \text{IG}(\nu, \omega),$$

where $GP(h, rk)$ stands for GP with mean function h and covariance function rk , and $\text{IG}(\nu, \omega)$ stands for an inverse gamma distribution with the density function

$$g(r) = \frac{1}{\Gamma(\nu)} \left(\frac{\omega}{r}\right)^{\nu+1} \frac{1}{\omega} \exp\left(-\frac{\omega}{r}\right),$$

and $\Gamma(\cdot)$ is the gamma function. Then, f follows an ETP $f \sim \text{ETP}(\nu, \omega, h, k)$, implying that for any collection of points $\mathbf{X}_n = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$, $\mathbf{x}_i \in \mathcal{X}$, we have

$$\mathbf{f}_n = f(\mathbf{X}_n) = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^T \sim \text{EMTD}(\nu, \omega, \mathbf{h}_n, \mathbf{K}_n),$$

where $\mathbf{f}_n \sim \text{EMTD}(\nu, \omega, \mathbf{h}_n, \mathbf{K}_n)$ means that \mathbf{f}_n has an extended multivari-

ate t -distribution (EMTD) with the density function,

$$p(z) = |2\pi\omega\mathbf{K}_n|^{-1/2} \frac{\Gamma(n/2 + \nu)}{\Gamma(\nu)} \left(1 + \frac{(z - \mathbf{h}_n)^T \mathbf{K}_n^{-1} (z - \mathbf{h}_n)}{2\omega} \right)^{-(n/2 + \nu)},$$

$\mathbf{h}_n = (h(\mathbf{x}_1), \dots, h(\mathbf{x}_n))^T$, $\mathbf{K}_n = (k_{ij})_{n \times n}$ and $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ for some mean function $h(\cdot) : \mathcal{X} \rightarrow R$ and kernel function $k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow R$.

It follows that at any collection of finite points ETP has an analytically representable EMTD density being similar to GP having multivariate normal density. Note that $E(\mathbf{f}_n) = \mathbf{h}_n$ is defined when $\nu > 1/2$ and $Cov(\mathbf{f}_n) = \omega\mathbf{K}_n/(\nu - 1)$ is defined when $\nu > 1$. When $\nu = \omega = \alpha/2$, \mathbf{f}_n becomes the multivariate t -distribution of Lange *et al.* (1989). When $\nu = \alpha/2$ and $\omega = \beta/2$, \mathbf{f}_n becomes the generalized multivariate t -distribution of Arellano-Valle and Bolfarine (1995). For $f \sim ETP(\nu, \omega, 0, k)$ it easily obtains that $E(f(\mathbf{x})) = 0$, $Var(f(\mathbf{x})) = \omega k(\mathbf{x}, \mathbf{x})/(\nu - 1)$, and

$$\begin{aligned} Skewness(f(\mathbf{x})) &= \frac{E(f^3(\mathbf{x}))}{(E(f^2(\mathbf{x})))^{3/2}} = 0, \\ Kurtosis(f(\mathbf{x})) &= \frac{E(f^4(\mathbf{x}))}{(E(f^2(\mathbf{x})))^2} = \frac{3}{\nu - 2} + 3 \geq 3 \text{ when } \nu > 2. \end{aligned}$$

Thus, we may say that the $ETP(\nu, \omega, 0, k)$ has a heavier tail than the $GP(0, k)$.

Proposition 1 *Let $f \sim ETP(\nu, \omega, h, k)$.*

- (i) *When $\omega/\nu \rightarrow \lambda$ as $\nu \rightarrow \infty$, we have $\lim_{\nu \rightarrow \infty} ETP(\nu, \omega, h, k) = GP(h, \lambda k)$.*
- (ii) *Let $\mathbf{Z} \in \mathcal{X}$ be a $p \times 1$ random vector such that $\mathbf{Z} \sim EMTD(\nu, \omega, \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$. For a linear system $f(\mathbf{x}) = \mathbf{x}^T \mathbf{Z}$ with $\mathbf{x} \in \mathcal{X}$, we have $f \sim ETP(\nu, \omega, h, k)$ with $h(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\mu}_z$ and $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \boldsymbol{\Sigma}_z \mathbf{x}_j$.*
- (iii) *Let $\mathbf{u} \in \mathcal{X}$ be a new data point and $\mathbf{k}_u = (k(\mathbf{u}, \mathbf{x}_1), \dots, k(\mathbf{u}, \mathbf{x}_n))^T$. Then, $f | \mathbf{f}_n \sim ETP(\nu^*, \omega^*, h^*, k^*)$ with $\nu^* = \nu + n/2$, $\omega^* = \omega + n/2$,*

$$\begin{aligned} h^*(\mathbf{u}) &= \mathbf{k}_u^T \mathbf{K}_n^{-1} (\mathbf{f}_n - \mathbf{h}_n) + h(\mathbf{u}), \\ k^*(\mathbf{u}, \mathbf{v}) &= \frac{2\omega + (\mathbf{f}_n - \mathbf{h}_n)^T \mathbf{K}_n^{-1} (\mathbf{f}_n - \mathbf{h}_n)}{2\omega + n} (k(\mathbf{u}, \mathbf{v}) - \mathbf{k}_u^T \mathbf{K}_n^{-1} \mathbf{k}_v), \end{aligned}$$

for $\mathbf{v} \in \mathcal{X}$.

Even if the mean and covariance functions of $f \sim ETP(\nu, \omega, h, k)$ cannot be defined when $\nu < 0.5$, from Proposition 1(iii), the mean and covariance functions of the conditional process $f|\mathbf{f}_n$ do always exist if $n \geq 2$. Also from Proposition 1(iii), the conditional process $ETP(\nu^*, \omega^*, h^*, k^*)$ converges to a GP, as either ν or n goes to ∞ . Thus, if the sample size n is large enough, the ETP behaves like a GP.

For a new point \mathbf{u} , we have $f(\mathbf{u})|\mathbf{f}_n \sim EMTD(\nu^*, \omega^*, h^*(\mathbf{u}), k^*(\mathbf{u}, \mathbf{u}))$, where

$$\begin{aligned} h^*(\mathbf{u}) &= E(f(\mathbf{u})|\mathbf{f}_n) = \mathbf{k}_{\mathbf{u}}^T \mathbf{K}_n^{-1}(\mathbf{f}_n - \mathbf{h}_n) + h(\mathbf{u}), \\ Var(f(\mathbf{u})|\mathbf{f}_n) &= \frac{\omega^*}{\nu^* - 1} k^*(\mathbf{u}, \mathbf{u}) = s \{k(\mathbf{u}, \mathbf{u}) - \mathbf{k}_{\mathbf{u}}^T \mathbf{K}_n^{-1} \mathbf{k}_{\mathbf{u}}\}, \end{aligned}$$

and $s = (2\omega + (\mathbf{f}_n - \mathbf{h}_n)^T \mathbf{K}_n^{-1}(\mathbf{f}_n - \mathbf{h}_n))/(2\nu + n - 2)$. Note that from Lemma 2(iv) $s = E(r|\mathbf{f}_n)$.

Under various combinations of ν and ω , the ETP generates various t -processes proposed in the literature. For example, $ETP(\alpha/2, \alpha/2 - 1, h, k)$ is the t -process of Shah *et al.* (2014). They showed that if covariance function Σ follows an inverse Wishart process with parameter $\alpha = 2\nu$ and kernel function k , and $f|\Sigma \sim GP(h, (\alpha - 2)\Sigma)$, then f has an extended t -process $ETP(\alpha/2, \alpha/2 - 1, h, k)$. $ETP(\alpha/2, \alpha/2, h, k)$ is the Student's t -process of Rasmussen and William (2006) and $ETP(\nu, 1/2, h, k)$ is that of Zhang and Yeung (2010).

3. eTPR models

Consider the process regression model (2)

$$y(\mathbf{x}) = f(\mathbf{x}) + \epsilon(\mathbf{x}), \quad \text{for } \mathbf{x} \in \mathcal{X}.$$

In this section we introduce an eTPR model, where f and ϵ have a joint ETP process,

$$\begin{pmatrix} f \\ \epsilon \end{pmatrix} \sim ETP\left(\nu, \omega, \begin{pmatrix} h \\ 0 \end{pmatrix}, \begin{pmatrix} k & 0 \\ 0 & \tilde{k} \end{pmatrix}\right), \quad (3)$$

kernel function $\tilde{k}(\mathbf{u}, \mathbf{v}) = \phi I(\mathbf{u} = \mathbf{v})$ and $I(\cdot)$ is an indicator function. The joint ETP above can be constructed hierarchically as

$$\begin{pmatrix} f \\ \epsilon \end{pmatrix} \Big| r \sim GP\left(\begin{pmatrix} h \\ 0 \end{pmatrix}, r \begin{pmatrix} k & 0 \\ 0 & \tilde{k} \end{pmatrix}\right) \quad \text{and} \quad r \sim \text{IG}(\nu, \omega),$$

and this implies that $f + \epsilon|r \sim GP(h, r(k + \tilde{k}))$ and $r \sim \text{IG}(\nu, \omega)$ to give $y \sim \text{ETP}(\nu, \omega, h, k + \tilde{k})$. Hence, additivity property of the GP and many other properties hold conditionally and marginally in the ETP. When $r = 1$, the eTPR model becomes a GPR model.

For observed data, this leads to a functional regression model

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where $y_i = y(\mathbf{x}_i)$ and $\epsilon_i = \epsilon(\mathbf{x}_i)$. Now it follows that

$$\begin{aligned} f(\mathbf{X}_n)|\mathbf{X}_n &\sim \text{EMTD}(\nu, \omega, \mathbf{h}_n, \mathbf{K}_n), \\ \mathbf{y}_n|f, \mathbf{X}_n &\sim \text{EMTD}(\nu, \omega, f(\mathbf{X}_n), \phi \mathbf{I}_n), \\ \mathbf{y}_n|\mathbf{X}_n &\sim \text{EMTD}(\nu, \omega, \mathbf{h}_n, \tilde{\Sigma}_n), \end{aligned}$$

where $\tilde{\Sigma}_n = \mathbf{K}_n + \phi \mathbf{I}_n$.

Consider a linear mixed model

$$y_i = \mathbf{w}_i^T \boldsymbol{\delta} + \mathbf{v}_i^T \mathbf{b} + \epsilon_i, \quad i = 1, \dots, n,$$

where \mathbf{w}_i is the design matrix for fixed effects $\boldsymbol{\delta}$, \mathbf{v}_i is the design matrix for random effect $\mathbf{b} \sim N(0, \theta \mathbf{I}_p)$ and $\epsilon_i \sim N(0, \phi)$ is a white noise. Suppose that $\mathbf{X}_n = (\mathbf{W}_n, \mathbf{V}_n)$, $f(\mathbf{X}_n) = \mathbf{W}_n^T \boldsymbol{\delta} + \mathbf{V}_n^T \mathbf{b}$, $\mathbf{h}_n = \mathbf{W}_n^T \boldsymbol{\delta}$ and $\mathbf{K}_n = \theta \mathbf{V}_n \mathbf{V}_n^T$ with $\mathbf{W}_n = (\mathbf{w}_1, \dots, \mathbf{w}_n)^T$ and $\mathbf{V}_n = (\mathbf{v}_1, \dots, \mathbf{v}_n)^T$. Then, the linear mixed model becomes the functional regression model with

$$\begin{aligned} f(\mathbf{X}_n)|\mathbf{X}_n &= \mathbf{W}_n^T \boldsymbol{\delta} + \mathbf{V}_n^T \mathbf{b} | \mathbf{X}_n \sim N(\mathbf{W}_n^T \boldsymbol{\delta}, \mathbf{K}_n), \\ \mathbf{y}_n|f, \mathbf{X}_n &= \mathbf{y}_n | \mathbf{b}, \mathbf{X}_n \sim N(\mathbf{W}_n^T \boldsymbol{\delta} + \mathbf{V}_n^T \mathbf{b}, \phi \mathbf{I}_n), \\ \mathbf{y}_n|\mathbf{X}_n &\sim N(\mathbf{W}_n^T \boldsymbol{\delta}, \tilde{\Sigma}_n). \end{aligned}$$

This shows that eTPR models extend the conventional normal linear mixed models to a nonlinear functional regression. In contrary to LOESS, this also shows that the eTPR method can produce a regression function, easily represented by a mathematical formula.

In the hierarchical construction of ETP, there is only one single random effect r , so that r is not estimable, confounded with parameters in covariance matrix. This means that ν and ω are not estimable. Following Lee and Nelder (2006), we set $\omega = \nu - 1$ because $\text{Var}(f) = \omega k / (\nu - 1) = k$ if $f \sim \text{ETP}(\nu, \omega, h, k)$. Thus, the variance does not depend upon ν as $\text{Var}(f) =$

k ; this is also true when $f \sim GP(h, k)$. By doing this way, the first two moments of GP and ETP have the same parametrization. Zellner (1976) also noted that ν cannot be estimated with a single realization of $\{(y_i, \mathbf{x}_i) : i = 1, 2, \dots, n\}$. In multivariate t-distribution, Lange *et al.* (1989) proposed to use $\nu = 2$. Zellner (1976) suggested that ν can be chosen according to investigator's knowledge of robustness of regression error distribution. As $\nu \rightarrow \infty$, ETP tends to GP. When robustness property is an important issue, a smaller ν is preferred. We tried various values for ν and find that $\nu = 1.05$ works well. From now on we set $\nu = 1.05$ to have $\omega = \nu - 1 = 0.05 > 0$. Furthermore, in functional regression models it is conventional to assume $h(\mathbf{u}) = 0$. Thus, without loss of generality we assume $h(\mathbf{u}) = 0$.

3.1. Parameter estimation for eTPR

So far we have assumed that the covariance kernel $k(\cdot, \cdot)$ is given. To fit the eTPR model, we need to choose $k(\cdot, \cdot)$. A way is to estimate the covariance kernel nonparametrically; see e.g. Hall *et al.* (2008). However, this method is very difficult to be applied to problems with multivariate covariates. Thus, we choose a covariance kernel from a function family such as a squared exponential kernel and Matérn class kernel. This paper employs a combination of a squared exponential and a non-stationary linear covariance kernel as follows,

$$k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}) = \theta_0 \exp \left(-\frac{1}{2} \sum_{l=1}^p \theta_{1l} (x_{i,l} - x_{j,l})^2 \right) + \sum_{l=1}^p \theta_{2l} x_{i,l} x_{j,l}, \quad (4)$$

where $\boldsymbol{\theta} = \{\theta_0, \theta_{1l}, \theta_{2l}, l = 1, \dots, p\}$ are a set of hyper-parameters. In (4), $1/\theta_{1l}$ measure the length scale of each input covariate, θ_0 known as scaling parameter which controls the vertical scale of variations of a typical function of the input, and θ_{2l} defines the scale of non-stationary linear trends. The small value of $1/\theta_{1l}$ means that the corresponding covariate may have great contribution in the covariance function. More about kernel function $k(\cdot, \cdot; \boldsymbol{\theta})$ can be seen in Rasmussen and William (2006) and Shi and Choi (2011).

Let $\boldsymbol{\beta} = (\phi, \boldsymbol{\theta})$ where ϕ is a parameter for $\epsilon(\mathbf{x})$ and $\boldsymbol{\theta}$ are those for $f(\mathbf{x})$. Here the joint density of $\mathbf{y}_n, f(\mathbf{X}_n) | \mathbf{X}_n$ is

$$p_{\boldsymbol{\beta}}(\mathbf{y}_n, f(\mathbf{X}_n) | \mathbf{X}_n) = p_{\phi}(\mathbf{y}_n | f, \mathbf{X}_n) p_{\boldsymbol{\theta}}(f(\mathbf{X}_n) | \mathbf{X}_n),$$

where $p_{\phi}(\mathbf{y}_n | f, \mathbf{X}_n)$ and $p_{\boldsymbol{\theta}}(f(\mathbf{X}_n) | \mathbf{X}_n)$ are density functions of EMTDs. Because $\mathbf{y}_n | \mathbf{X}_n \sim EMTD(\nu, \nu - 1, 0, \tilde{\boldsymbol{\Sigma}}_n)$, the maximum likelihood (ML)

estimator $\hat{\beta}$ for β can be obtained by solving

$$\frac{\partial \log p_{\beta}(\mathbf{y}_n | \mathbf{X}_n)}{\partial \beta} = \frac{1}{2} Tr \left(\left(s_1 \alpha \alpha^T - \tilde{\Sigma}_n^{-1} \right) \frac{\partial \tilde{\Sigma}_n}{\partial \beta} \right) = 0,$$

where $\alpha = \tilde{\Sigma}_n^{-1} \mathbf{y}_n$, $s_1 = (n + 2\nu) / (2(\nu - 1) + \mathbf{y}_n^T \tilde{\Sigma}_n^{-1} \mathbf{y}_n)$, and

$$p_{\beta}(\mathbf{y}_n | \mathbf{X}_n) = |2\pi(\nu - 1) \tilde{\Sigma}_n|^{-1/2} \frac{\Gamma(n/2 + \nu)}{\Gamma(\nu)} \left(1 + \frac{\mathbf{y}_n^T \tilde{\Sigma}_n^{-1} \mathbf{y}_n}{2(\nu - 1)} \right)^{-(n/2 + \nu)}. \quad (5)$$

The score equations for GPR models above are ML estimating equations in linear mixed models with $\nu = \infty$ and $s_1 = 1$. Thus, a little modification of existing BLUP procedures gives a parameter estimation for eTPR models.

3.2. Predictive distribution

Since

$$\left(\begin{array}{c} f(\mathbf{X}_n) \\ \mathbf{y}_n \end{array} \right) \bigg| \mathbf{X}_n \sim EMTD \left(\nu, \nu - 1, 0, \left(\begin{array}{cc} \mathbf{K}_n & \mathbf{K}_n \\ \mathbf{K}_n & \tilde{\Sigma}_n \end{array} \right) \right),$$

from Lemma 2(iii) we have $f(\mathbf{X}_n) | \mathcal{D}_n = \{\mathbf{X}_n, \mathbf{y}_n\} \sim EMTD(n/2 + \nu, n/2 + \nu - 1, \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, with

$$\begin{aligned} \boldsymbol{\mu}_n &= E(f(\mathbf{X}_n) | \mathcal{D}_n) = \mathbf{K}_n \tilde{\Sigma}_n^{-1} \mathbf{y}_n, \\ \boldsymbol{\Sigma}_n &= Cov(f(\mathbf{X}_n) | \mathcal{D}_n) = s_0 \phi \mathbf{K}_n \tilde{\Sigma}_n^{-1}, \\ s_0 &= E(r | \mathcal{D}_n) = \frac{\mathbf{y}_n^T \tilde{\Sigma}_n^{-1} \mathbf{y}_n + 2(\nu - 1)}{n + 2(\nu - 1)}. \end{aligned}$$

Thus, given β , $E(f(\mathbf{X}_n) | \mathcal{D}_n)$ is linear in \mathbf{y}_n , i.e. the BLUP for $f(\mathbf{X}_n)$, which is an extension of the BLUP in linear mixed models to eTPR models. This BLUP has a form independent of ν , so that it is the BLUP for GPR models. However, the conditional variance depends upon ν , except when $r = 1$, i.e. $s_0 = 1$ under GPR models. Thus, the BLUPs for the eTPR and GPR models have a common form, but have different predictors and their variance estimations because of different parameter estimations ($s_0 \neq 1$ and $s_1 \neq 1$). Furthermore, all quantities necessary to compute s_0 and s_1 are available during implementing BLUP procedures.

For a given new data point \mathbf{u} , we have

$$\begin{pmatrix} \mathbf{y}_n \\ f(\mathbf{u}) \end{pmatrix} \Big| \mathbf{X}_n \sim EMTD \left(\nu, \nu - 1, 0, \begin{pmatrix} \tilde{\Sigma}_n & \mathbf{k}\mathbf{u} \\ \mathbf{k}\mathbf{u}^T & k(\mathbf{u}, \mathbf{u}) \end{pmatrix} \right).$$

By Lemma 2(iii), the predictive distribution $p(f(\mathbf{u})|\mathcal{D}_n)$ is $EMTD(n/2 + \nu, n/2 + \nu - 1, \mu_n^*, \sigma_n^*)$, where

$$\mu_n^* = E(f(\mathbf{u})|\mathcal{D}_n) = \mathbf{k}\mathbf{u}^T \tilde{\Sigma}_n^{-1} \mathbf{y}_n, \quad (6)$$

$$\sigma_n^* = Var(f(\mathbf{u})|\mathcal{D}_n) = s_0 \left(k(\mathbf{u}, \mathbf{u}) - \mathbf{k}\mathbf{u}^T \tilde{\Sigma}_n^{-1} \mathbf{k}\mathbf{u} \right). \quad (7)$$

Furthermore, from Proposition 1(iii), $f|\mathcal{D}_n \sim ETP(n/2 + \nu, n/2 + \nu - 1, h^*, k^*)$, where $h^*(\mathbf{u}) = \mu_n^*$ and $k^*(\mathbf{u}, \mathbf{v}) = s_0 \left(k(\mathbf{u}, \mathbf{v}) - \mathbf{k}\mathbf{u}^T \tilde{\Sigma}_n^{-1} \mathbf{k}\mathbf{v} \right)$. From Lemma 2(iii), we also have $y(\mathbf{u})|\mathcal{D}_n \sim EMTD(n/2 + \nu, n/2 + \nu - 1, \mu_n^*, \sigma_n^* + s_0\phi)$ with $E(y(\mathbf{u})|\mathcal{D}_n) = \mu_n^*$ and $Var(y(\mathbf{u})|\mathcal{D}_n) = \sigma_n^* + s_0\phi$. Consequently, this conditional predictive process can be used to construct prediction $\hat{y}(\mathbf{u}) = E(y(\mathbf{u})|\mathcal{D}_n) = E(f(\mathbf{u})|\mathcal{D}_n)$ of the unobserved response $y(\mathbf{u})$ at $\mathbf{x} = \mathbf{u}$ and its standard error can be formed using the predictive variance, given by $\sigma_n^* + s_0\phi$, and the proof is in Appendix. The predictive variance for $\hat{f}(\mathbf{u})$ in (7) differs from that for $\hat{y}(\mathbf{u})$.

The prediction of $f(\mathbf{X}_n)$ and $f(\mathbf{u})$ discussed above is the best unbiased predictions under eTPR models, and so is under GPR models. However, their standard errors (variance estimators) differ. Note that

$$s_0 = \frac{\mathbf{y}_n^T \tilde{\Sigma}_n^{-1} \mathbf{y}_n + 2(\nu - 1)}{n + 2(\nu - 1)} = \frac{(\mathbf{y}_n - \hat{\mathbf{f}}_n)^T \tilde{\Sigma}_n (\mathbf{y}_n - \hat{\mathbf{f}}_n) / \phi^2 + 2(\nu - 1)}{n + 2(\nu - 1)},$$

where $\hat{\mathbf{f}}_n$ is the BLUP for $f(\mathbf{X}_n)$. Thus, the standard error estimate of the BLUP under the eTPR model increases if the model does not fit the responses \mathbf{y}_n well while that under the GPR model does not depend upon the model fit.

Random-effect models consist with three objects, namely the data \mathcal{D}_n , unobservables (random effects) and parameters (fixed unknowns) β . For inferences of such models, Lee and Nelder (1996) proposed the use of the h-likelihood. Lee and Kim (2015) showed that inferences about unobservables allow both Bayesian and frequentist interpretations. In this paper, we see that the eTPR model is an extension of random-effect models. Thus, we may

view the functional regression model (2) either as a Bayesian model, where a GP or an ETP as a prior, or as a frequentist model where a latent process such as GP and ETP is used to fit unknown function $f_0()$ in a functional space (Chapter 9, Lee *et al.*, 2006). With the predictive distribution above, we may form both Bayesian credible and frequentist confidence intervals. Estimation procedures in Section 3.1 can be viewed as an empirical Bayesian method with a uniform prior on β . In frequentist (or Bayesian) approach, (5) is a marginal likelihood for fixed (or hyper) parameters.

3.3. Robust properties

Let $\hat{f}_T(\mathbf{u}) = \hat{\mu}_n^* = \mu_n^*|_{\beta=\hat{\beta}}$ and $V_T = \hat{\sigma}_n^* = \sigma_n^*|_{\beta=\hat{\beta}}$ be the BLUP for $f(\mathbf{u})$ and its variance estimate, respectively, under the eTPR model. And let $\hat{f}_G(\mathbf{u})$ and V_G be those under the GPR model with $s_0 = 1$. Let $M_T = (\hat{f}_T(\mathbf{u}) - f_0(\mathbf{u}))/\sqrt{V_T}$ and $M_G = (\hat{f}_G(\mathbf{u}) - f_0(\mathbf{u}))/\sqrt{V_G}$ be two student t-type statistics for a null hypothesis $f(\mathbf{u}) = f_0(\mathbf{u})$. Under a bounded kernel function, if $y_i \rightarrow \infty$ for some i , $M_G \rightarrow \infty$, while M_T remains bounded. Therefore, M_T for eTPR is more robust against outliers in output space compared to that for GPR. This property still holds for ML estimators.

Proposition 2 *If kernel function $k(\mathbf{u}, \mathbf{v}; \theta)$ is bounded, continuous and differentiable on θ , then the ML estimator $\hat{\beta}$ from the eTPR has bound influence function, while that from the GPR does not.*

3.4. Information Consistency

Let $p_{\phi_0}(\mathbf{y}_n|f_0, \mathbf{X}_n)$ be the density function to generate the data \mathbf{y}_n given \mathbf{X}_n under the true model (1), where f_0 is the true underlying function of f . Let $p_{\theta}(f)$ be a measure of random process f on space $\mathcal{F} = \{f(\cdot) : \mathcal{X} \rightarrow R\}$. Let

$$p_{\phi, \theta}(\mathbf{y}_n|\mathbf{X}_n) = \int_{\mathcal{F}} p_{\phi}(\mathbf{y}_n|f, \mathbf{X}_n) dp_{\theta}(f),$$

be the density function to generate the data \mathbf{y}_n given \mathbf{X}_n under the assumed eTPR model (3). Thus, the assumed model (3) is not the same as the true underlying model (1). Here ϕ is the common in both models and ϕ_0 is the true value of ϕ . Let $p_{\phi_0, \hat{\theta}}(\mathbf{y}_n|\mathbf{X}_n)$ be the estimated density function under the eTPR model. Denote $D[p_1, p_2] = \int (\log p_1 - \log p_2) dp_1$ by the Kullback-Leibler distance between two densities p_1 and p_2 . Then, we have the following proposition.

Proposition 3 *Under the appropriate conditions in Appendix, we have*

$$\frac{1}{n}E_{\mathbf{X}_n}(D[p_{\phi_0}(\mathbf{y}_n|f_0, \mathbf{X}_n), p_{\phi_0, \hat{\theta}}(\mathbf{y}_n|\mathbf{X}_n)]) \longrightarrow 0, \text{ as } n \rightarrow \infty,$$

where the expectation is taken over the distribution of \mathbf{X}_n .

From Proposition 3, the Kullback-Leibler distance between two density functions for $\mathbf{y}_n|\mathbf{X}_n$ from the true and the assumed models becomes zero, asymptotically. Let $\mathbf{y}_i = (y_1, \dots, y_i)^T$ and $\mathbf{X}_i = (\mathbf{x}_1, \dots, \mathbf{x}_i)^T$, $i = 1, \dots, n$. In Appendix, we show that

$$p_{\phi_0, \theta}(\mathbf{y}_n|\mathbf{X}_n) = \prod_{i=1}^n p_{\phi_0, \theta}(y_i|\mathbf{X}_i, \mathbf{y}_{i-1}), \quad (8)$$

where

$$p_{\phi_0, \theta}(y_i|\mathbf{X}_i, \mathbf{y}_{i-1}) = \int_{\mathcal{F}} p_{\phi_0}(y_i|f, \mathbf{X}_i, \mathbf{y}_{i-1}) dp_{\theta}(f|\mathbf{X}_i, \mathbf{y}_{i-1}),$$

$$p_{\theta}(f|\mathbf{X}_i, \mathbf{y}_{i-1}) = \frac{p_{\phi_0}(\mathbf{y}_{i-1}|f, \mathbf{X}_{i-1})}{\int_{\mathcal{F}} p_{\phi_0}(\mathbf{y}_{i-1}|f', \mathbf{X}_{i-1}) dp_{\theta}(f')}.$$

Under the true model (1), similarly to (8), we have

$$p_{\phi_0}(\mathbf{y}_n|f_0, \mathbf{X}_n) = \prod_{i=1}^n p_{\phi_0}(y_i|f_0, \mathbf{X}_i, \mathbf{y}_{i-1}).$$

Seeger *et al.* (2008) called $p_{\phi_0}(y_i|f_0, \mathbf{X}_i, \mathbf{y}_{i-1})$ and $p_{\phi_0, \hat{\theta}}(y_i|\mathbf{X}_i, \mathbf{y}_{i-1})$ Bayesian prediction strategies. We can show that

$$D[p_{\phi_0}(\mathbf{y}_n|f_0, \mathbf{X}_n), p_{\phi_0, \hat{\theta}}(\mathbf{y}_n|\mathbf{X}_n)] = \int \sum_{i=1}^n Q(y_i|\mathbf{X}_i, \mathbf{y}_{i-1}) p_{\phi_0}(\mathbf{y}_n|f_0, \mathbf{X}_n) d\mathbf{y}_n,$$

where $Q(y_i|\mathbf{X}_i, \mathbf{y}_{i-1}) = \log\{p_{\phi_0}(y_i|f_0, \mathbf{X}_i, \mathbf{y}_{i-1})/p_{\phi_0, \hat{\theta}}(y_i|\mathbf{X}_i, \mathbf{y}_{i-1})\}$ is a loss function and $\sum_{i=1}^n Q(y_i|\mathbf{X}_i, \mathbf{y}_{i-1})$ is called cumulative loss. Under the GPR model, Seeger *et al.* (2008) and Wang and Shi (2014) proved Proposition 3, interpreted it as the average of cumulative loss $\sum_{i=1}^n Q(y_i|\mathbf{X}_i, \mathbf{y}_{i-1})/n$ tending to zero asymptotically, and called it the information consistency. In this paper, we show this property for the robust BLUPs. Consequently, the frequentist BLUP procedure is consistent with the Bayesian strategy in terms of average risk over an ETP prior.

4. Numerical studies

4.1. Simulation studies

We use simulation studies to evaluate performance of the BLUP procedures from the eTPR model (3). For GPR and eTPR models, we use

- GPR: $f \sim GP(0, k)$ and $\epsilon_i \sim N(0, \phi)$;
- eTPR: $f \sim ETP(\nu, \nu - 1, 0, k)$ and $\epsilon \sim ETP(\nu, \nu - 1, 0, \tilde{k})$;

where kernel function k is given in (4) and $\tilde{k}(u, v) = \phi I(u = v)$. Results are based on 500 simulation data.

Selective shrinkage

When some sparse data points are far away from the dense data points, predictions of the sparse ones from the eTPR method are more heavily regularized than those from the LOESS and the GPR methods. To generate data, from the process model (2) we assume f follows a GP with mean 0 and the kernel function (4), and error term follows a normal distribution with mean 0 and variance ϕ , denoted by $N(0, \phi)$. We set $\beta = (\phi, \theta_0, \theta_{11}, \theta_{21}) = (0.1, 0.05, 10, 0.05)$. In Figure 1, 95% prediction confidence bounds are computed as $\hat{f}(u) \pm 1.96\sqrt{Var(\hat{f}(u))}$. At sparse data point, from Figure 1 we see that the eTPR method has selective shrinkage of Wauthier and Jordan (2010) and gives a wider interval.

We compare prediction performance from the LOESS, GPR and eTPR methods in Table 1, where $n = 10$ and the data point 2.0 is added with an extra error from either $N(0, \sigma^2)$ or σt_2 with $\sigma^2 = 1, 2, 3$ and 4. Testing data points are evenly spaced from interval $(0, 2.0)$, denoted by $\{x_j^* : j = 1, \dots, m\}$ with $m = 30$. Prediction performance of the test data points is measured with mean squared error $MSE = \sum_{j=1}^m \hat{f}(x_j^*)^2 / m$. Table 1 shows that robust BLUPs from the eTPR model have the smallest MSE among the three methods: LOESS, GPR and eTPR. The improvement is greater with t error.

Instead of random disturbance, a constant disturbance is added to the last data point 2 in training data. Let $y_n^* = y_n + \delta$ where $\delta = -2, -1, 0, 1$ and 2. Then predicted values $\hat{y}(u)$ at data points $u = 0, 1.0, 1.5, 1.8$ and 2.0 are calculated by the LOESS, GPR and eTPR. In each data point, Figure 2 shows an average value of predictions from 500 simulated data, where the true function is 0, and dashed, dotted and solid lines respectively represent average

Table 1: Mean squared errors of prediction results and their standard deviation (in parentheses) by the LOESS, GPR and eTPR methods with $\beta = (0.1, 0.05, 0.05, 10)$.

	error	σ^2	LOESS	GPR	eTPR
t	Normal	1	0.238(0.236)	0.204(0.255)	0.167(0.220)
		2	0.330(0.363)	0.305(0.397)	0.235(0.333)
		3	0.421(0.493)	0.406(0.533)	0.300(0.440)
		4	0.513(0.623)	0.507(0.669)	0.357(0.539)
	t	1	0.735(3.183)	0.721(3.142)	0.390(1.298)
		2	1.054(3.063)	0.974(2.106)	0.518(1.086)
		3	1.498(4.575)	1.372(3.053)	0.818(3.462)
		4	1.646(3.269)	1.614(3.165)	0.859(1.998)

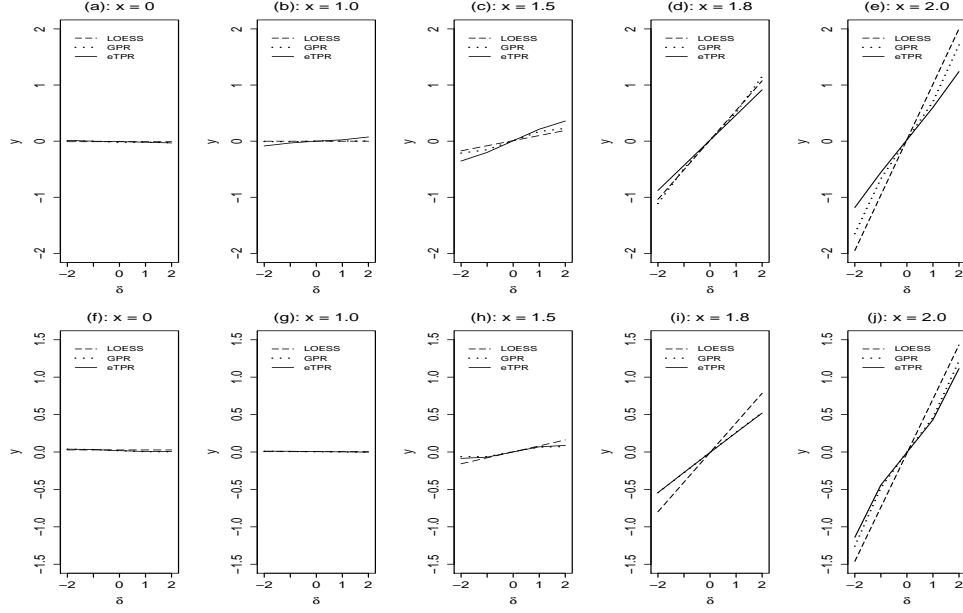


Figure 2: Predicted values at the data points $x \in \{0.0, 1.0, 1.5, 1.8, 2.0\}$ with constant disturbance at the point 2.0 for sample sizes 10 (sub-figures a-g) and 50 (sub-figures f-j), where dashed, dotted and solid lines respectively represent predictions from the LOESS, GPR and eTPR methods.

prediction from the LOESS, GPR and eTPR methods. In a small sample with $n = 10$, Figure 2(a-e) shows that the predictions from the LOESS and the GPR methods tend to be shrunken more at dense region $0 \leq x \leq 1.5$, while those from the eTPR method are shrunken heavily at sparse region $1.5 < x \leq 2.0$. For moderate sample size $n = 50$, it follows from Figure 2(f-j) that the eTPR behaves like the GPR, and the eTPR shrinks data points 1.8 and 2.0 more heavily than the LOESS.

Robust property against outliers in output space

We generate the data y_i under five process models as follows:

- (1) $f \sim GP(h, k)$, $\epsilon \sim N(0, \phi)$ and $\beta = (0.1, 0.01, 10, 0.01) = \beta_1$;
 - (2) $f \sim GP(h, k)$, $\epsilon \sim N(0, \phi)$ and $\beta = (0.2, 0.2, 10, 0.1) = \beta_2$;
 - (3) $f \sim GP(h, k)$, $\epsilon \sim \phi t_2$ and $\beta = \beta_1$;
 - (4) $f \sim GP(h, k)$, $\epsilon \sim \phi t_2$ and $\beta = \beta_2$;
 - (5) $f \sim ETP(2, 2, h, k)$, $\epsilon \sim ETP(2, 2, 0, \tilde{k})$ and $\beta = (0.1, 0.02, 10, 0.02) = \beta_3$,
- where $h(x) = \cos(x)$ or $\cos(2x)$ for $x \in (0, 3)$. Let S be a set of 40 points evenly spaced in the interval $(0, 3)$. We randomly take $n = 10$ data points from S as the training data set, and the rest as test data set. Values of mean squared error, $MSE = \sum_{j=1}^m (\hat{f}(x_i^*) - h(x_i^*))^2 / m$ for test data points $\{x_i^*, i = 1, \dots, m\}$, are computed by using the LOESS, GPR and eTPR methods. Table 2 shows the MSEs for Cases (1) and (2), where the data are generated from GPR models. We can see that all three methods work similarly. Now we consider cases for model misspecifications and/or the presence of outliers. For Cases (1), (2) and (5), one data point is randomly selected from the training data set and is added with a t_1 error to study robustness of the proposed methods. Now Cases (1) and (2) have outliers, Cases (3) and (4) have non-normal errors and Case (5) has both. We see from Table 3 that the eTPR method gives BLUPs with much smaller MSE than the LOESS and GPR methods.

We also study robustness of BLUPs from the eTPR model with multi-variate covariates. We consider $h_1(\mathbf{x}) = 0.5x_1|x_1|^{1/3} - 3\cos(x_2) + \log(x_3)$ and $h_2(\mathbf{x}) = 0.2x_1^3 + \sin(x_2) + 0.2\exp(x_3)$ with $\mathbf{x} = (x_1, x_2, x_3)^T$. In this case, parameter β is $(\phi, \theta_0, \theta_1, \theta_2)$ with $\theta_1 = (\theta_{11}, \theta_{12}, \theta_{13})$ and $\theta_2 = (\theta_{21}, \theta_{22}, \theta_{23})$. To generate the data, we follow the previous five process models, but $\beta_1 = (0.1, 0.01, 10, 10, 10, 0.01, 0.01, 0.01)$, $\beta_2 = (0.2, 0.05, 10, 10, 10, 0.05, 0.05, 0.05)$, and $\beta_3 = (0.1, 0.02, 10, 10, 10, 0.02, 0.02, 0.02)$. Let S_1 , S_2 and S_3 be sets of 80 points evenly spaced in the intervals $(-2, 2)$, $(0, 3)$ and $(1, 2)$, respectively. We take $n = 30$ random points as training data and the remaining $m = 50$ points as test data. For Cases (1), (2) and (5), two data points

are randomly selected from the training data set and are added with two independent t_1 errors. Table 4 presents MSE results. Again, BLUPs from the eTPR method is better than those from the LOESS and GPR methods. As the number of covariates increases, the LOESS has the worst MSE.

Table 2: Mean squared errors of prediction results and their standard deviation (in parentheses) by the LOESS, GPR and eTPR methods with function $h(x) = \cos(x)$ and $\cos(2x)$ and Cases (1) and (2) of data generation.

function	Model	LOESS	GPR	eTPR
cos(x)	(1)	0.450(0.492)	0.455(0.486)	0.450(0.481)
	(2)	0.587(0.620)	0.556(0.554)	0.549(0.555)
cos(2x)	(1)	0.459(0.499)	0.514(0.498)	0.502(0.486)
	(2)	0.595(0.629)	0.627(0.560)	0.643(0.603)

Table 3: Mean squared errors of prediction results and their standard deviation (in parentheses) by the LOESS, GPR and eTPR methods with function $h(x) = \cos(x)$ and $\cos(2x)$.

function	Model	LOESS	GPR	eTPR
cos(x)	(1)	1.765(7.831)	0.812(1.634)	0.569(0.743)
	(2)	1.938(7.778)	0.901(1.612)	0.666(1.030)
	(3)	0.887(4.014)	0.641(1.117)	0.611(1.027)
	(4)	1.137(1.461)	0.886(1.880)	0.808(1.495)
	(5)	2.130(8.511)	0.838(3.956)	0.355(0.458)
cos(2x)	(1)	1.771(7.827)	0.912(1.484)	0.665(0.777)
	(2)	1.943(7.775)	0.974(1.447)	0.720(0.751)
	(3)	0.891(3.998)	0.741(1.132)	0.684(1.121)
	(4)	1.139(1.453)	0.996(1.867)	0.900(1.653)
	(5)	2.030(8.119)	0.752(3.086)	0.447(0.465)

4.2. Real examples

The eTPR model (3) is applied to three data sets. Executive function research data come from the study in children with Hemiplegic Cerebral

Table 4: Mean squared errors of prediction results and their standard deviation (in parentheses) by the LOESS, GPR and eTPR methods with multivariate mean functions $h(\mathbf{x}) = h_1(\mathbf{x})$ and $h_2(\mathbf{x})$.

$h(\mathbf{x})$	Model	LOESS	GPR	eTPR
$h_1(\mathbf{x})$	(1)	1.086(5.000)	0.468(1.669)	0.453(1.761)
	(2)	1.237(5.001)	0.774(1.771)	0.768(1.728)
	(3)	0.654(0.425)	0.272(1.350)	0.207(0.445)
	(4)	0.884(0.829)	0.738(2.684)	0.632(0.941)
	(5)	1.212(4.255)	0.761(2.247)	0.640(1.452)
$h_2(\mathbf{x})$	(1)	1.120(4.471)	0.490(2.088)	0.306(0.837)
	(2)	1.262(4.489)	0.813(2.100)	0.615(0.882)
	(3)	0.753(0.437)	0.289(0.580)	0.234(0.297)
	(4)	0.987(0.872)	0.745(1.075)	0.634(0.653)
	(5)	1.312(4.405)	0.718(1.781)	0.528(0.788)

Palsy consisting of 84 girls and 57 boys from primary and secondary schools. These students were subdivided into two groups: the action video game players group (AVGPs) (56%) and the non action video game players group (NAVGP) (44%). In this study, Big/Little Circle (BLC) mean correct latency is investigated as age of children: for more details of this data set, see Xu *et al.* (2015). Before applying the proposed methods, we take logarithm of Big/Little Circle (BLC) mean correct latency. Figure 3 presents prediction curves for 2 groups: AVGPs and NAVGP, where circles represent observed data points, and solid line, dashed line and dotted line stand for predictions from the GPR, eTPR and LOESS methods, respectively. We can see prediction curves from the LOESS and eTPR methods are more smooth than those from the GPR method.

Whistler snowfall data contain daily snowfall amounts in Whistler for the years 2010 and 2011, and can be downloaded at <http://www.climate.weatheroffice.gc.ca>. Response for snow data is logarithm of (daily snowfall amount+1) and covariate is time. From Figure 3, we can see that predicted curve from the LOESS is the most smooth, while that from the GPR is the least smooth.

For spatial interpolation data, rainfall measurements at 467 locations were recorded in Switzerland on 8 May 1986, and can be found at <http://www.ai->

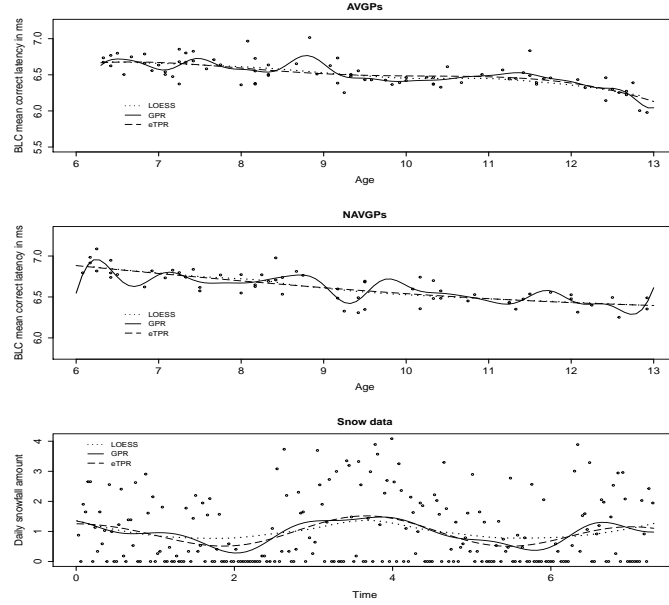


Figure 3: Prediction curves from the LOESS, GPR and eTPR methods for 2 groups of the BLC data and snow data, where circles represent data points, and solid, dotted and dashed lines stand for predictions from the GPR, LOESS and eTPR, respectively.

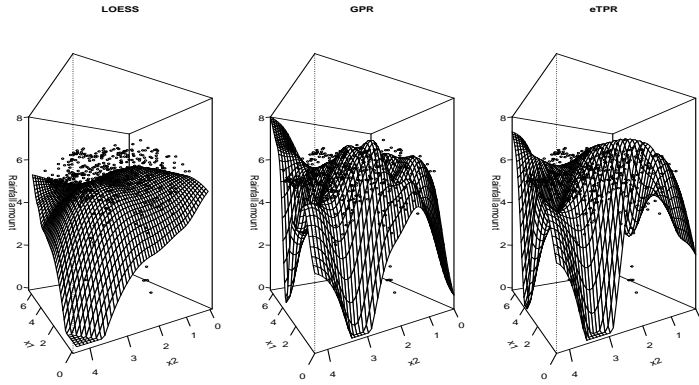


Figure 4: Prediction surfaces from the LOESS, GPR and eTPR methods for spatial data, where circles represent data points.

geostats.org under SIC97. Spatial interpolation data has response, logarithms of (rainfall amount+1), and two covariates for coordinates of location. Prediction surfaces of spatial data are presented in Figure 4. We can see again that the LOESS surface is the most smooth while the GPR one is the least smooth.

We randomly select 80% observation as training data and compute prediction errors for the remaining data points (i.e. the test data). This procedure is repeated 500 times. Table 5 presents mean prediction errors of these 3 data sets. We can see that the LOESS is the best in BLC-AVGPs, while it is the worst in the snow and spatial data particularly for the latter which includes multivariate predictors. Overall, the eTPR is the best in prediction.

Table 5: Prediction errors and their standard deviation (in parentheses) for the 3 real data sets by the LOESS, GPR and eTPR methods.

Data	LOESS	GPR	eTPR
BLC-AVGPs	0.022(0.010)	0.031(0.015)	0.024(0.010)
BLC-NAVGPs	0.017(0.006)	0.025(0.033)	0.016(0.006)
Snow	1.133(0.097)	1.117(0.102)	1.116(0.101)
Spatial	0.524(0.121)	0.210(0.086)	0.204(0.089)

5. Concluding remarks

Advantages of a GPR model include that it offers a nonparametric regression model for data with multi-dimensional covariates, the specification of covariance kernel enables to accommodate a wide class of nonlinear regression functions, and it can be applied to analyze many different types of data including functional data. In this paper, we extended the GPR model to the eTPR model. The latter inherits almost all the good features for the GPR, and additionally it provides robust BLUP procedures in the presence of outliers in both input and output spaces. Numerical studies show that the eTPR is overall the best in prediction among the methods considered.

Appendix

Let Σ be an $n \times n$ symmetric and positive definite matrix, $\boldsymbol{\mu} \in R^n$, $\nu > 0$ and $\omega > 0$. In this paper, $\mathbf{Z} \sim EMTD(\nu, \omega, \boldsymbol{\mu}, \Sigma)$ means that a random

vector $\mathbf{Z} \in R^n$ has the density function,

$$p(\mathbf{z}) = |2\pi\omega\mathbf{\Sigma}|^{-1/2} \frac{\Gamma(n/2 + \nu)}{\Gamma(\nu)} \left(1 + \frac{(\mathbf{z} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu})}{2\omega} \right)^{-(n/2 + \nu)},$$

where $\Gamma(\cdot)$ is the gamma function.

We may construct an EMTD via a double hierarchical generalized linear model (Lee and Nelder, 2006) as follows:

Lemma 1 *If*

$$\mathbf{Z}|r \sim N(\boldsymbol{\mu}, r\mathbf{\Sigma}), \quad r \sim \text{IG}(\nu, \omega),$$

where $\text{IG}(\nu, \omega)$ stands for an inverse gamma distribution with the density function

$$g(r) = \frac{1}{\Gamma(\nu)} \left(\frac{\omega}{r} \right)^{\nu+1} \frac{1}{\omega} \exp\left(-\frac{\omega}{r}\right),$$

then, marginally $\mathbf{Z} \sim \text{EMTD}(\nu, \omega, \boldsymbol{\mu}, \mathbf{\Sigma})$.

Proof: From the construction of \mathbf{Z} , we have

$$\begin{aligned} p(\mathbf{z}) &= \int_0^\infty p(\mathbf{z}|r)g(r)dr \\ &= \int_0^\infty |2\pi r\mathbf{\Sigma}|^{-1/2} \exp\left(-\frac{(\mathbf{z} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu})}{2r}\right) \frac{1}{\Gamma(\nu)} \left(\frac{\omega}{r} \right)^{\nu+1} \frac{1}{\omega} \exp\left(-\omega/r\right) dr \\ &= \int_0^\infty |2\pi\omega\mathbf{\Sigma}|^{-1/2} \frac{1}{\Gamma(\nu)} \left(\frac{\omega}{r} \right)^{n/2 + \nu - 1} \exp\left(-\frac{2\omega + (\mathbf{z} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu})}{2r}\right) d\frac{\omega}{r} \\ &= |2\pi\omega\mathbf{\Sigma}|^{-1/2} \frac{\Gamma(n/2 + \nu)}{\Gamma(\nu)} \left(1 + \frac{(\mathbf{z} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu})}{2\omega} \right)^{-(n/2 + \nu)}, \end{aligned}$$

which is the density function of EMTD.‡

Properties of EMTD are as follows.

Lemma 2 *Let $\mathbf{Z} \sim \text{EMTD}(\nu, \omega, \boldsymbol{\mu}, \mathbf{\Sigma})$.*

- (i) *If $\omega/\nu \rightarrow \lambda > 0$ as $\nu \rightarrow \infty$, then $\lim_{\nu \rightarrow \infty} \text{EMTD}(\nu, \omega, \boldsymbol{\mu}, \mathbf{\Sigma}) = N(\boldsymbol{\mu}, \lambda\mathbf{\Sigma})$.*
- (ii) *For any matrix $\mathbf{A} \in R^{l \times n}$ with $\text{rank } l \leq n$, $\mathbf{AZ} \sim \text{EMTD}(\nu, \omega, \mathbf{A}\boldsymbol{\mu}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T)$.*

- (iii) Let \mathbf{Z} be partitioned as $(\mathbf{Z}_1^T, \mathbf{Z}_2^T)^T$ with lengths n_1 and $n_2 = n - n_1$, and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have the corresponding partitions as $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{pmatrix}$. Then,

$$\begin{aligned}\mathbf{Z}_1 &\sim \text{EMTD}(\nu, \omega, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \\ \mathbf{Z}_2 | \mathbf{Z}_1 = \mathbf{z}_1 &\sim \text{EMTD}(\nu^*, \omega^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*),\end{aligned}$$

with $\nu^* = n_1/2 + \nu$, $\omega^* = n_1/2 + \omega$, $\boldsymbol{\mu}^* = \boldsymbol{\Sigma}_{12}^T \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}^* = (2\omega + (\mathbf{z}_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1)) \boldsymbol{\Sigma}_{22 \cdot 1} / (2\omega + n_1)$, and $\boldsymbol{\Sigma}_{22 \cdot 1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$. This gives $E(\mathbf{Z}_2 | \mathbf{Z}_1) = \boldsymbol{\mu}^*$ and $\text{Cov}(\mathbf{Z}_2 | \mathbf{Z}_1) = \omega^* \boldsymbol{\Sigma}^* / (\nu^* - 1)$.

- (iv) Let r be a random effect in Proposition 1. Then, $r | \mathbf{Z} \sim \text{IG}(\tilde{\nu}, \tilde{\omega})$ with $\tilde{\nu} = n/2 + \nu$, $\tilde{\omega} = \omega + (\mathbf{Z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{\mu})/2$ and

$$\begin{aligned}E(r | \mathbf{Z}) &= \frac{2\omega + (\mathbf{Z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{\mu})}{n + 2\nu - 2}, \\ \text{Var}(r | \mathbf{Z}) &= \frac{(2\omega + (\mathbf{Z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{\mu}))^2}{(n + 2\nu - 2)^2(n/2 + \nu - 2)}.\end{aligned}$$

Proof: The conclusions (i), (ii) and $\mathbf{Z}_1 \sim \text{EMTD}(\nu, \omega, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ in (iii) are easily obtained by the definition of EMTD and Lemma 1. Now we only prove that $\mathbf{Z}_2 | \mathbf{Z}_1 \sim \text{EMTD}(\nu^*, \omega^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. Let $a_1 = (\mathbf{z}_1 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{z}_1 - \boldsymbol{\mu}_1)$ and $a_2 = (\mathbf{z}_2 - \boldsymbol{\mu}^*)^T \boldsymbol{\Sigma}^{*-1}(\mathbf{z}_2 - \boldsymbol{\mu}^*)$, then $a_1 + a_2 = (\mathbf{z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu})$. We have

$$\begin{aligned}p(\mathbf{z}_2 | \mathbf{z}_1) &= \frac{p(\mathbf{z})}{p(\mathbf{z}_1)} \\ &= \frac{|2\pi\omega\boldsymbol{\Sigma}|^{-1/2} \frac{\Gamma(n/2+\nu)}{\Gamma(\nu)} \left(1 + \frac{a_1+a_2}{2\omega}\right)^{-(n/2+\nu)}}{|2\pi\omega\boldsymbol{\Sigma}_{11}|^{-1/2} \frac{\Gamma(n_1/2+\nu)}{\Gamma(\nu)} \left(1 + \frac{a_1}{2\omega}\right)^{-(n_1/2+\nu)}} \propto \left(1 + \frac{a_2}{2\omega + a_1}\right)^{-(n/2+\nu)},\end{aligned}$$

which indicates $\mathbf{Z}_2 | \mathbf{Z}_1 \sim \text{EMTD}(\nu^*, \omega^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$.

By combining definitions of IG and EMTD, we have

$$\begin{aligned}
p(r|\mathbf{Z}) &= \frac{p(\mathbf{Z}|r)g(r)}{p(\mathbf{Z})} \\
&= \frac{1}{\Gamma(n/2 + \nu)} \frac{1}{\omega + (\mathbf{Z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \boldsymbol{\mu})/2} \left(\frac{\omega + (\mathbf{Z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \boldsymbol{\mu})/2}{r} \right)^{n/2 + \nu + 1} \\
&\quad \exp \left(-\frac{\omega + (\mathbf{Z} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Z} - \boldsymbol{\mu})/2}{r} \right),
\end{aligned}$$

which indicates (iv) holds in this Lemma. \sharp

Proof of Proposition 1: Proposition 1 can be easily proved by using Lemma 2, so omitted here. \sharp

Marginal likelihood derivatives:

We know that $\mathbf{y}_n | \mathbf{X}_n \sim EMTD(\nu, \nu - 1, 0, \tilde{\boldsymbol{\Sigma}}_n)$. For given ν , the marginal log-likelihood of $\boldsymbol{\beta}$ is

$$\begin{aligned}
l(\boldsymbol{\beta}; \nu) &= -\frac{n}{2} \log(2\pi(\nu - 1)) - \frac{1}{2} \log |\tilde{\boldsymbol{\Sigma}}_n| - \left(\frac{n}{2} + \nu \right) \log \left(1 + \frac{S}{2(\nu - 1)} \right) \\
&\quad + \log(\Gamma(\frac{n}{2} + \nu)) - \log(\Gamma(\nu)),
\end{aligned}$$

where $S = \mathbf{y}_n^T \tilde{\boldsymbol{\Sigma}}_n^{-1} \mathbf{y}_n$. The derivative with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial l(\boldsymbol{\beta}; \nu, (\nu - 1))}{\partial \boldsymbol{\beta}} = \frac{1}{2} Tr \left(\left(\frac{n + 2\nu}{2(\nu - 1) + S} \boldsymbol{\alpha} \boldsymbol{\alpha}^T - \tilde{\boldsymbol{\Sigma}}_n^{-1} \right) \frac{\partial \tilde{\boldsymbol{\Sigma}}_n}{\partial \boldsymbol{\beta}} \right), \quad (9)$$

where $\boldsymbol{\alpha} = \tilde{\boldsymbol{\Sigma}}_n^{-1} \mathbf{y}_n$.

Estimates of parameters $\boldsymbol{\beta}$ can be learned by using gradient based methods. And variances of the estimates can be estimated by computing the second derivatives of $l(\boldsymbol{\beta}; \nu)$ on $\boldsymbol{\beta}$ as follows,

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\beta}; \nu)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} &= \frac{1}{2} Tr \left(\left(\frac{n + 2\nu}{2(\nu - 1) + S} \boldsymbol{\alpha} \boldsymbol{\alpha}^T - \tilde{\boldsymbol{\Sigma}}_n^{-1} \right) \left(\frac{\partial^2 \tilde{\boldsymbol{\Sigma}}_n}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} - \frac{\partial \tilde{\boldsymbol{\Sigma}}_n}{\partial \boldsymbol{\beta}} \tilde{\boldsymbol{\Sigma}}_n^{-1} \frac{\partial \tilde{\boldsymbol{\Sigma}}_n}{\partial \boldsymbol{\beta}} \right) \right) \\
&\quad - \frac{1}{2} Tr \left(\frac{n + 2\nu}{2(\nu - 1) + S} \boldsymbol{\alpha} \boldsymbol{\alpha}^T \frac{\partial \tilde{\boldsymbol{\Sigma}}_n}{\partial \boldsymbol{\beta}} \tilde{\boldsymbol{\Sigma}}_n^{-1} \frac{\partial \tilde{\boldsymbol{\Sigma}}_n}{\partial \boldsymbol{\beta}} \right) + \frac{1}{2} \frac{n + 2\nu}{(2(\nu - 1) + S)^2} \left\{ Tr \left(\boldsymbol{\alpha} \boldsymbol{\alpha}^T \frac{\partial \tilde{\boldsymbol{\Sigma}}_n}{\partial \boldsymbol{\beta}} \right) \right\}^2.
\end{aligned}$$

Variance of prediction value $\hat{y}(\mathbf{u})$:

From the hierarchical sampling method described in Lemma 1, we have

$$\begin{pmatrix} f \\ \epsilon \end{pmatrix} \Big| r \sim GP \left(\nu, (\nu - 1), \begin{pmatrix} h \\ 0 \end{pmatrix}, \begin{pmatrix} rk & 0 \\ 0 & r\tilde{k} \end{pmatrix} \right), \quad r \sim \text{IG}(\nu, (\nu - 1)),$$

which suggests that conditional on r , $\mathbf{y}_n | f, \mathbf{X}_n \sim N(f(\mathbf{X}_n), r\phi \mathbf{I}_n)$ and marginal distribution of $\mathbf{y}_n | \mathbf{X}_n \sim N(\mathbf{h}_n, r\tilde{\Sigma}_n)$. For given r , it follows from the GPR model that $E(\hat{f}(\mathbf{u}) | r, \mathcal{D}_n) = \mathbf{k}_{\mathbf{u}}^T \tilde{\Sigma}_n^{-1} (\mathbf{y}_n - \mathbf{h}_n) + h(\mathbf{u})$ and $\text{Var}(\hat{f}(\mathbf{u}) | r, \mathcal{D}_n) = r(k(\mathbf{u}, \mathbf{u}) - \mathbf{k}_{\mathbf{u}}^T \tilde{\Sigma}_n^{-1} \mathbf{k}_{\mathbf{u}} + \phi)$. Consequently, we have

$$\begin{aligned} \text{Var}(\hat{f}(\mathbf{u}) | \mathcal{D}_n) &= E((\hat{f}(\mathbf{u}))^2 | \mathcal{D}_n) - (E(\hat{f}(\mathbf{u}) | \mathcal{D}_n))^2 \\ &= E_r((\text{Var}(\hat{f}(\mathbf{u}) | r, \mathcal{D}_n) + (E(\hat{f}(\mathbf{u}) | r, \mathcal{D}_n))^2) | \mathcal{D}_n) - (E(\hat{f}(\mathbf{u}) | \mathcal{D}_n))^2 \\ &= E_r(\text{Var}(\hat{f}(\mathbf{u}) | r, \mathcal{D}_n) | \mathcal{D}_n) + (E(\hat{f}(\mathbf{u}) | \mathcal{D}_n))^2 - (E(\hat{f}(\mathbf{u}) | \mathcal{D}_n))^2 \\ &= s_0 \left(k(\mathbf{u}, \mathbf{u}) - \mathbf{k}_{\mathbf{u}}^T \tilde{\Sigma}_n^{-1} \mathbf{k}_{\mathbf{u}} + \phi \right), \end{aligned}$$

where $s_0 = (2\nu - 2 + (\mathbf{y}_n - \mathbf{h}_n)^T \tilde{\Sigma}_n^{-1} (\mathbf{y}_n - \mathbf{h}_n)) / (2\nu + n - 2)$.

Proof of Proposition 2: From (9), the score functions of β based on the eTPR model is

$$s_T(\beta; \mathbf{y}_n) = \frac{1}{2} Tr \left(\left(\frac{n + 2\nu}{2(\nu - 1) + \mathbf{y}_n^T \tilde{\Sigma}_n^{-1} \mathbf{y}_n} \tilde{\Sigma}_n^{-1} \mathbf{y}_n \mathbf{y}_n^T \tilde{\Sigma}_n^{-1} - \tilde{\Sigma}_n^{-1} \right) \frac{\partial \tilde{\Sigma}_n}{\partial \beta} \right).$$

The term $(n + 2\nu) / (2(\nu - 1) + \mathbf{y}_n^T \tilde{\Sigma}_n^{-1} \mathbf{y}_n)$ in $s_T(\beta; \mathbf{y}_n)$ plays an important role in estimating parameter β . For example, when some observations of responses have very large value or tend to infinity (outliers), the score $s_T(\beta; \mathbf{y}_n)$ based on the eTPR model does not tend to infinity.

Let $\mathbf{T}(F_n) = \mathbf{T}_n(y_1, \dots, y_n)$ be an estimate of β , where F_n is the empirical distribution of $\{y_1, \dots, y_n\}$ and T is a functional on some subset of all distributions. Influence function of T at F (Hampel *et al.*, 1986) is defined as

$$IF(y; T, F) = \lim_{t \rightarrow 0} \frac{T((1 - t)F + t\delta_y) - T(F)}{t},$$

where δ_y put mass 1 on point y and 0 on others.

For given parameter ν , following Hampel *et al.* (1986) estimator $\hat{\beta}$ of β has the influence function

$$IF(y; \hat{\beta}, F) = - \left(E \left(\frac{\partial^2 l(\beta; \nu, (\nu - 1))}{\partial \beta \partial \beta^T} \right) \right)^{-1} s_T(\beta; y).$$

Note that the matrix $\partial^2 l(\beta; \nu) / \partial \beta \partial \beta^T$ is bounded according to \mathbf{y}_n , which indicates that the influence function of $\hat{\beta}$ is bounded under the eTPR model. Similarly, we can obtain that the score function under the GPR model is unbound, which leads to unbound influence function of parameter estimate.‡

Proof of the equation (8):

From sequential bayesian prediction strategy and Bayes' Theorem, we have

$$\begin{aligned} \prod_{i=1}^n p_{\phi_0, \theta_0}(y_i | \mathbf{X}_i, \mathbf{y}_{i-1}) &= p_{\phi_0, \theta_0}(y_1 | \mathbf{X}_1) \prod_{i=2}^n \int_{\mathcal{F}} p_{\phi_0}(y_i | f, \mathbf{X}_i, \mathbf{y}_{i-1}) dp(f | \mathbf{X}_i, \mathbf{y}_{i-1}) \\ &= p_{\phi_0, \theta_0}(y_1 | \mathbf{X}_1) \prod_{i=2}^n \int_{\mathcal{F}} \frac{p_{\phi_0}(\mathbf{y}_i | f, \mathbf{X}_i) dp_{\theta_0}(f)}{\int_{\mathcal{F}} p_{\phi_0}(\mathbf{y}_{i-1} | f', \mathbf{X}_{i-1}) dp_{\theta_0}(f')} \\ &= \int_{\mathcal{F}} p_{\phi_0}(\mathbf{y}_n | f, \mathbf{X}_n) dp_{\theta_0}(f) = p_{\phi_0, \theta_0}(\mathbf{y}_n | \mathbf{X}_n), \end{aligned}$$

which shows that the equation (8) holds.‡

Lemma 3 Suppose $\mathbf{y}_n = \{y_1, \dots, y_n\}$ are generated from the eTPR model (3) with the mean function $h(\mathbf{x}) = 0$, and covariance kernel function k is bounded and continuous in parameter θ . It also assumes that the estimate $\hat{\beta}$ almost surely converges to β as $n \rightarrow \infty$. Then for a positive constant c , and any $\varepsilon > 0$, when n is large enough, we have

$$\begin{aligned} &\frac{1}{n} (-\log p_{\phi_0, \hat{\theta}}(\mathbf{y}_n | \mathbf{X}_n) + \log p_{\phi_0}(\mathbf{y}_n | f_0, \mathbf{X}_n)) \\ &\leq \frac{1}{n} \left\{ \frac{1}{2} \log |\mathbf{I}_n + \phi_0^{-1} \mathbf{K}_n| + \frac{s^2 + 2(\nu - 1)}{2(n + 2\nu - 2)} (\|f_0\|_k^2 + c) + c \right\} + \varepsilon, \end{aligned}$$

where $\mathbf{K}_n = (k(x_i, x_j))_{n \times n}$, $s^2 = (\mathbf{y}_n - f_0(\mathbf{X}_n))^T (\mathbf{y}_n - f_0(\mathbf{X}_n)) / \phi_0$, \mathbf{I}_n is the $n \times n$ identity matrix, and $\|f_0\|_k$ is the reproducing kernel Hilbert space norm of f_0 associated with kernel function $k(\cdot, \cdot; \theta)$.

Proof: From Proposition 1, it follows that there exists a variable $r \sim IG(\nu, (\nu - 1))$ with density function $g(r)$, conditional on r we have

$$\begin{pmatrix} f \\ \epsilon \end{pmatrix} | r \sim GP \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} rk & 0 \\ 0 & r\tilde{k} \end{pmatrix} \right),$$

where $GP(h, k)$ stands for Gaussian process with mean function h and covariance function k . Then conditional on r , the extended t-process regression model (2) becomes Gaussian process regression model

$$y(\mathbf{x}) = \tilde{f}(\mathbf{x}) + \tilde{\epsilon}(\mathbf{x}), \quad (10)$$

where $\tilde{f} = f|r \sim GP(0, rk(\cdot, \cdot; \theta))$, $\tilde{\epsilon}|r \sim GP(0, r\tilde{k}(\cdot, \cdot; \phi_0))$, and \tilde{f} and error term $\tilde{\epsilon}$ are independent. Denoted \tilde{p} by probability density computation conditional on r . Based on the model (10), let

$$\begin{aligned} p_G(\mathbf{y}_n | r, \mathbf{X}_n) &= \int_{\mathcal{F}} \tilde{p}_{\phi_0}(\mathbf{y}_n | \tilde{f}, \mathbf{X}_n) d\tilde{p}_n(\tilde{f}), \\ p_0(\mathbf{y}_n | r, \mathbf{X}_n) &= \tilde{p}_{\phi_0}(\mathbf{y}_n | f_0, \mathbf{X}_n), \end{aligned}$$

where \tilde{p}_n is the induced measure from Gaussian process $GP(0, rk(\cdot, \cdot; \hat{\theta}))$.

We know that random effect r is independent of covariates \mathbf{X}_n . Then it easily shows that

$$p_{\phi_0, \hat{\theta}}(\mathbf{y}_n | \mathbf{X}_n) = \int p_G(\mathbf{y}_n | r, \mathbf{X}_n) g(r) dr, \quad (11)$$

$$p_{\phi_0}(\mathbf{y}_n | f_0, \mathbf{X}_n) = \int p_0(\mathbf{y}_n | r, \mathbf{X}_n) g(r) dr. \quad (12)$$

Suppose that for any given r , we have

$$\begin{aligned} & -\log p_G(\mathbf{y}_n | r, \mathbf{X}_n) + \log p_0(\mathbf{y}_n | r, \mathbf{X}_n) \\ & \leq \frac{1}{2} \log |\mathbf{I}_n + \phi_0^{-1} \mathbf{K}_n| + \frac{r}{2} (\|f_0\|_k^2 + c) + c + n\varepsilon, \end{aligned} \quad (13)$$

which indicates

$$\begin{aligned} & -\log \int p_G(\mathbf{y}_n | r, \mathbf{X}_n) g(r) dr \leq \frac{1}{2} \log |\mathbf{I}_n + \phi_0^{-1} \mathbf{K}_n| + c + n\varepsilon \\ & -\log \int p_0(\mathbf{y}_n | r, \mathbf{X}_n) \exp\left\{-\left(\frac{r}{2} (\|f_0\|_k^2 + c)\right)\right\} g(r) dr. \end{aligned} \quad (14)$$

By simple computation, it follows that

$$\begin{aligned} & \int p_0(\mathbf{y}_n|r, \mathbf{X}_n) \exp\left\{-\left(\frac{r}{2}(\|f_0\|_k^2 + c)\right)\right\} g(r) dr \\ &= \int p_0(\mathbf{y}_n|r, \mathbf{X}_n) g(r) dr \int \exp\left\{-\left(\frac{r}{2}(\|f_0\|_k^2 + c)\right)\right\} \tilde{g}(r) dr, \end{aligned} \quad (15)$$

where $\tilde{g}(r)$ is the density function of $IG(\nu + n/2, (\nu - 1) + s^2/2)$. From (11), (12), (14) and (15), we have

$$\begin{aligned} & -\log p_{\phi_0, \hat{\theta}}(\mathbf{y}_n|\mathbf{X}_n) + \log p_{\phi_0}(\mathbf{y}_n|f_0, \mathbf{X}_n) \\ & \leq \frac{1}{2} \log |\mathbf{I}_n + \phi_0^{-1} \mathbf{K}_n| + c - \log \int \exp\left\{-\left(\frac{r}{2}(\|f_0\|_k^2 + c)\right)\right\} \tilde{g}(r) dr \\ & \leq \frac{1}{2} \log |\mathbf{I}_n + \phi_0^{-1} \mathbf{K}_n| + c + \frac{\|f_0\|_k^2 + c}{2} \int r \tilde{g}(r) dr \\ & = \frac{1}{2} \log |\mathbf{I}_n + \phi_0^{-1} \mathbf{K}_n| + \frac{s^2 + 2(\nu - 1)}{2(n + 2\nu - 2)} (\|f_0\|_k^2 + c) + c + n\varepsilon, \end{aligned}$$

which shows that Lemma 3 holds.

Now let us prove the inequality (13). Since the proof of (13) is similar to those of Theorem 1 in Seeger *et al.* (2008) and Lemma 1 in Wang and Shi (2014), here we summarily present the procedure of the proof, details please see in Seeger *et al.* (2008) and Wang and Shi (2014). Let \mathcal{H} be the reproducing kernel Hilbert space (RKHS) associated with covariance function $k(\cdot, \cdot; \boldsymbol{\theta})$, and $\mathcal{H}_n = \{\tilde{f}(\cdot) : \tilde{f}(\cdot) = \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i; \boldsymbol{\theta}), \text{ for any } \alpha_i \in R\}$. From the Representer Theorem (see Lemma 2 in Seeger *et al.*, 2008), it is sufficient to prove (13) for the true underlying function $\tilde{f}_0 = f_0|r \in \mathcal{H}_n$. Then for given r , f_0 can be written as

$$f_0(\cdot) = r \sum_{i=1}^n \alpha_i k(\mathbf{x}, \mathbf{x}_i; \boldsymbol{\theta}) \doteq r K(\cdot) \boldsymbol{\alpha},$$

where $K(\cdot) = (k(\mathbf{x}, \mathbf{x}_1; \boldsymbol{\theta}), \dots, k(\mathbf{x}, \mathbf{x}_n; \boldsymbol{\theta}))$ and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^T$.

By Fenchel-Legendre duality relationship, we have

$$-\log p_G(\mathbf{y}_n|r, \mathbf{X}_n) \leq E_Q(-\log \tilde{p}(\mathbf{y}_n|\tilde{f})) + D[Q, P], \quad (16)$$

where P is a measure induced by $GP(0, rk(\cdot, \cdot; \hat{\boldsymbol{\theta}}_n))$, and Q is the posterior distribution of \tilde{f} from a GP model with prior $GP(0, rk(\cdot, \cdot; \boldsymbol{\theta}))$ and Gaussian

likelihood term $\prod_{i=1}^n N(\hat{y}_i | \tilde{f}(\mathbf{x}_i), r\phi_0)$, where $\hat{\mathbf{y}} = (\hat{y}_1, \dots, \hat{y}_n)^T = r(\mathbf{K}_n + \phi_0 \mathbf{I}_n)\boldsymbol{\alpha}$ and $\mathbf{K}_n = (k(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta}))$. Then we have $E_Q(\tilde{f}) = f_0$. Let $\mathbf{B} = \mathbf{I}_n + \phi_0^{-1} \mathbf{K}_n$, then we have

$$D[Q, P] = \frac{1}{2} \left\{ -\log |\hat{\mathbf{K}}_n^{-1} \mathbf{K}_n| + \log |\mathbf{B}| + \text{tr}(\hat{\mathbf{K}}_n^{-1} \mathbf{K}_n \mathbf{B}^{-1}) + r \|f_0\|_k^2 + r \boldsymbol{\alpha} \mathbf{K}_n (\hat{\mathbf{K}}_n^{-1} \mathbf{K}_n - \mathbf{I}_n) \boldsymbol{\alpha} - n \right\}, \quad (17)$$

$$\begin{aligned} E_Q(-\log \tilde{p}(\mathbf{y}_n | \tilde{f})) &\leq -\log \tilde{p}(\mathbf{y}_n | f_0) + \frac{1}{2} \phi_0^{-1} \text{tr}(\mathbf{K}_n \mathbf{B}^{-1}) \\ &= -\log p_0(\mathbf{y}_n | r, \mathbf{X}_n) + \frac{1}{2} \phi_0^{-1} \text{tr}(\mathbf{K}_n \mathbf{B}^{-1}), \end{aligned} \quad (18)$$

where $\hat{\mathbf{K}}_n = (k(\mathbf{x}_i, \mathbf{x}_j; \hat{\boldsymbol{\theta}}))$.

Hence, it follows from (16), (17) and (18) that

$$\begin{aligned} &-\log p_G(\mathbf{y}_n | r, \mathbf{X}_n) + \log p_0(\mathbf{y}_n | r, \mathbf{X}_n) \\ &\leq \frac{1}{2} \left\{ -\log |\hat{\mathbf{K}}_n^{-1} \mathbf{K}_n| + \log |\mathbf{B}| + \text{tr}((\hat{\mathbf{K}}_n^{-1} \mathbf{K}_n + \phi_0^{-1} \mathbf{K}_n) \mathbf{B}^{-1}) + r \|f_0\|_k^2 \right. \\ &\quad \left. + r \boldsymbol{\alpha} \mathbf{K}_n (\hat{\mathbf{K}}_n^{-1} \mathbf{K}_n - \mathbf{I}_n) \boldsymbol{\alpha} - n \right\}. \end{aligned} \quad (19)$$

Since the covariance function is bounded and continuous in $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}$, we have $\hat{\mathbf{K}}_n^{-1} \mathbf{K}_n - \mathbf{I}_n \rightarrow 0$ as $n \rightarrow \infty$. Hence, there exist positive constants c and ε such that for n large enough

$$\begin{aligned} -\log |\hat{\mathbf{K}}_n^{-1} \mathbf{K}_n| &< c, \quad \boldsymbol{\alpha} \mathbf{K}_n (\hat{\mathbf{K}}_n^{-1} \mathbf{K}_n - \mathbf{I}_n) \boldsymbol{\alpha} < c, \\ \text{tr}(\hat{\mathbf{K}}_n^{-1} \mathbf{K}_n \mathbf{B}^{-1}) &< \text{tr}((\mathbf{I}_n + \varepsilon \mathbf{K}_n) \mathbf{B}^{-1}). \end{aligned} \quad (20)$$

Plugging (20) in (19), we have the inequality (13). \sharp

For proof of Proposition 3, we need condition

(A) $\|f_0\|_k$ is bounded and $E_{\mathbf{X}_n}(\log |\mathbf{I}_n + \phi_0^{-1} \mathbf{K}_n|) = o(n)$.

Proof of Proposition 3: It easily shows that $s^2 = (\mathbf{y}_n - f_0(\mathbf{X}_n))^T (\mathbf{y}_n - f_0(\mathbf{X}_n)) / \phi_0 = O(n)$. Under conditions in Lemma 3, and condition (A), it follows from Lemma 3 that

$$\frac{1}{n} E_{\mathbf{X}_n}(D[p_{\phi_0}(\mathbf{y}_n | f_0, \mathbf{X}_n), p_{\phi_0, \hat{\boldsymbol{\theta}}}(\mathbf{y}_n | \mathbf{X}_n)]) \longrightarrow 0, \text{ as } n \rightarrow \infty.$$

That proves Proposition 3. \sharp

References

1. Archambeau, C. and Bach, F. (2010), Multiple Gaussian Process Models. *Advances in Neural Information Processing Systems*.
2. Arellano-Valle, R. B. and Bolfarine, H. (1995). On some characterization of the t-distribution. *Statistics & Probability Letters*, 25, 79 - 85.
3. Cleveland, W.S., Devlin, S.J. (1988), Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* 83: 596-610.
4. Dutta, S. and Mondal, D. (2015), An h-likelihood method for spatial mixed linear models based on intrinsic auto-regressions. *J. R. Statist. Soc. B* 77: 699-726.
5. Hall, P., Müller, H.-G., and Yao, F. (2008), Modelling Sparse Generalized Longitudinal Observations with Latent Gaussian Processes, *Journal of Royal Statistical Society, Ser. B*, 70, 703-723.
6. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), Robust Statistics: The Approach Based on Influence Functions, Wiley.
7. Lange, K.L., Little, R. J.A. and Taylor J. M.G. (1989), Robust statistical modelling using the t distribution, *Journal of the American Statistical Association*, 84, 881-896.
8. Lee, Y. and Kim, G. (2015). H-likelihood predictive intervals for unobservables, *International Statistical Review*, DOI: 10.1111/insr.12115.
9. Lee, Y. and Nelder, J.A. (1996). Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society B*, 58, 619-678.
10. Lee, Y. and Nelder, J.A. (2006). Double hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society: C (Applied Statistics)*, 55, 139-185.
11. Lee, Y., Nelder, J.A. and Pawitan, Y. (2006). Generalized Linear Models with Random Effects, Unified Analysis via H-likelihood. Chapman & Hall/CRC.
12. Ma, R. and Jorgensen, B. (2007). Nested generalized linear mixed models: an orthodox best linear unbiased predictor approach. *Journal of the Royal Statistical Society B*, 69, 625-641.
13. Rasmussen, C. E. and Williams, C. K. I. (2006), Gaussian Processes for Machine Learning. Cambridge, Massachusetts: The MIT Press.

14. Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science* 6, 15-51.
15. Seeger M. W., Kakade S. M. and Foster D. P. (2008), Information Consistency of Nonparametric Gaussian Process Methods, *IEEE Transactions on Information Theory*, 54, 2376-2382.
16. Shah A., Wilson A.G. and Ghahramani Z. (2014), Student-t processes as alternatives to Gaussian processes. *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 877-885.
17. Shi, J. Q. and Choi, T. (2011), *Gaussian Process Regression Analysis for Functional Data*, London: Chapman and Hall/CRC.
18. Wang, B. and Shi, J.Q. (2014), Generalized Gaussian process regression model for non-Gaussian functional data. *Journal of the American Statistical Association*, 109, 1123-1133.
19. Wauthier, F. L. and Jordan, M. I. (2010). Heavy-tailed process priors for selective shrinkage. In *Advances in Neural Information Processing Systems*, 2406-2414.
20. Xu, P, Lee, Y. and Shi, J. Q. (2015) Automatic Detection of Significant Areas for Functional Data with Directional Error Control. *arXiv:1504.08164*.
21. Xu, Z., Yan, F. and Qi, Y. (2011), Sparse Matrix-Variate t Process Blockmodel. *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 543-548.
22. Yu S., Tresp V. and Yu K. (2007), Robust multi-task learning with t -process. *Proceedings of the 24th International Conference on Machine Learning*, 1103-1110.
23. Zellener, A. (1976), Bayesian and non-Bayesian analysis of the regression model with multivariate student- t error terms, *Journal of the American Statistical Association*, 71, 400-405.
24. Zhang, Y. and Yeung, D.Y. (2010), Multi-task learning using generalized t process. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 964-971.
25. Zhou, X. and Stephens, M. (2012), Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* 44: 821-824.